Санкт-Петербургский государственный университет
Математическое обеспечение и администрирование информационных систем
Кафедра информационно-аналитических систем

Скачкаускайте Анна Гинтарасовна

Анализ методов оценки качества получаемых данных, при проведении исследования территории, планируемой к застройке объектами альтернативной энергетики

Выпускная квалификационная работа

Научный руководитель:
к. ф.-м. н., доцент Михайлова Е. Г.
Рецензент:
к.т.н. Жук С.В. “Digital Design”

Санкт-Петербург
2017

SAINT-PETERSBURG STATE UNIVERSITY
Software and Administration of Information Systems
Sub-Department of Analytical Information Systems

Anna Skachkauskayte

Analysis of methods for evaluation of data quality obtained in investigating territory planned for construction by alternative energy facilities

Graduation Project

Scientific supervisor:
Associate Professor, Ph.D Mikhailova E. G
Reviewer:
Ph.D. in Engineering Zhuk S.V. “Digital Design”

Saint-Petersburg
2017

Оглавление

Введение

Альтернативная энергетика — совокупность перспективных способов получения, передачи и использования энергии, которые распространены не так широко, как традиционные, однако представляют интерес из-за выгодности их использования при, как правило, низком риске причинения вреда окружающей среде.[1] К альтернативным источникам энергии относятся возобновляемые источники – энергия солнца, ветра, геотермальная, океаническая, энергия биомассы, термоядерная энергия и другие источники[12].

В настоящее время данная сфера активно развивается во множестве направлений в техническом, экономическом, социальном и информационном, а также исследования в данной сфере поддерживаются и продвигаются правительствами многих стран, в том числе и России. И если для нашей страны альтернативная энергетика является довольно молодой областью, то в Китае в 2012 году был сформирован Национальный центр исследования ВИЭ. Несмотря на это, в некоторых областях использования ВИЭ Россия имеет крупные научные результаты, соответствующие мировому уровню.

Разработки в области альтернативной энергетики актуальны не только для крупных промышленных разработок, но для инноваций, применяемых в быту. Примером этого может послужить разработка солнечных зарядных устройств для различных гаджетов, которая получила широкое распространение в современном мире.

Перед застройкой определенной территории объектами альтернативной энергетики необходимо провести длительное исследование местности с

[1] Wikipedia URL: https://ru.wikipedia.org/wiki/Альтернативная_энергетика (Дата обращения: 22.04.2017)

учетом множества факторов (климатические данные, экономические и пр.), при этом накапливается огромное количество разнообразных данных, которые должны быть правильно структурированы, проанализированы и обладать определенным качеством.

Постановка задачи

Цель работы – исследование существующих алгоритмов и методов для проведения оценки качества получаемых данных, при проведении прединвестиционного исследования территории, планируемой к застройке объектами альтернативной энергетики.

Актуальность

Разработки в сфере альтернативной энергетики актуальны по целому ряду причин, среди которых можно выделить следующие[13-17]:

- примерно ⅔ территории России находится вне сетей централизованного энергоснабжения
- несмотря на статус газовой державы, лишь 50% городских и около 35% сельских населенных пунктов в России газифицированы[2]
- многие регионы страны нуждаются в завозе топлива или же поставке энергии (т. е. они энергодефицитны)

В настоящее время особенностью использования ВИЭ является высокая стоимость получаемой энергии (в сравнении с энергией, получаемой на традиционных электростанциях). Несмотря на это в России существуют крупные зоны, где по многим условиям приоритетно развитие возобновляемых источников энергии. Это, например, зоны децентрализованного энергоснабжения с низкой плотностью населения

[2] URL: http://www.energosovet.ru/bul_stat.php?idd=210/ (Дата обращения: 24.04.2017)

или города и места массового отдыха и лечения населения со сложной экологической обстановкой.[3]

На данный момент Российским правительством разработан план для проектирования и строительства объектов ВИЭ до 2035 года,а также выделено 250 млрд. рублей на строительство объектов ВИЭ до 2020 года.

Не только крупные частные российские компании, одна из которых «Газпром», инвестируют в строительство ВИЭ на территории Российской Федерации, но также это направление интересно и для иностранных инвесторов, среди которых китайские компании, инвестирующие при поддержке министерства Китая.

Важно отметить, что данная область для нашей страны является достаточно молодой, поэтому в ней существует большое количество нерешенных задач,которые подлежат автоматизации в ближайшее время.

[3] URL: http://ars24.pro/usefulmats/vozobnovlyaemye-istochniki-energii-vie/ (Дата обращения: 16.04.2017)

Первая глава. Обзор литературы

При исследовании местности производится анализ множества климатических данных, представленных в виде временных рядов. В качестве характеристик могут быть использованы следующие комбинации:

- температура/время замера
- влажность/время замера
- сила и направление ветра/время замера
- сила соляризации/время замера
- и т.д.

Но при передаче данных в силу различных обстоятельств (технический сбой оборудования, погодные помехи при передаче сигнала и т.д.). Это создает серьезные проблемы для проведения точных исследований. Поэтому для их устранения разработан ряд специальных методов и алгоритмов, повышающих качество данных.

1 Понятие качества данных

Качеством данных называют характеристику, показывающую степень пригодности данных к анализу. Приведение данных в соответствие с критериями качества является важнейшей задачей анализа данных и называется предобработкой.

Качество данных исключительно важно для анализа, прогнозирования. Даже если каждая из систем, поставляющих данные для проекта, содержит

лишь небольшой процент "плохих" данных, то при их объединении этот процент растет по экспоненциальному закону.

1.1 Кластерный анализ

Кластерный анализ или кластеризация – задача группировки объектов таким способом, что объекты, находящиеся в одной группе (названной кластером), в некотором смысле более подобны друг другу, чем объекты в других группах (кластерах).[11] Это основная задача анализа данных и общий метод для анализа статистических данных, используемого во многих сферах, включая машинное обучение, распознавание образов, анализ изображений, информационный поиск, биоинформатику, сжатие данных и компьютерную графику.

Задача кластерного анализа может быть решена различными алгоритмами, которые значительно различаются понятием того, что составляет кластер и как эффективно их определить.

Кластеризация может быть определена как многоцелевая задача оптимизации. Надлежащий алгоритм кластеризации и установки параметров зависят от конкретного набора данных и надлежащего использования результатов. Кластерный анализ как таковой не автоматическая задача, а итеративный процесс изучения знаний или интерактивная многоцелевая оптимизация. Часто необходимо использовать предварительную обработку данных, пока результат не достигает желаемых значений.

1.2 Классификация

В машинном обучении и статистике – задача идентификации, которой из ряда категорий принадлежит новое наблюдение, на основе набора данных,

содержащих наблюдения (или экземпляры), чье членство в некоторой категории известно. Примером является распределение электронной почты на классы "спама" и "неспама" или определение диагноза пациенту, по характеристиками пациента (пол, кровяное давление, присутствие или отсутствие определенных заболеваний, и т.д.). Классификация – пример распознавания образов.

1.3 Регрессия

В статистическом моделировании регрессионный анализ – статистический процесс для оценки отношений между переменными. Он включает много методов для моделирования и анализа нескольких переменных, при этом фокус находится на отношении между зависимой переменной и одной или нескольких независимых переменных. Обычно, регрессионный анализ оценивает условное ожидание зависимой переменной, при заданных независимых переменных, т.е. среднее значение зависимой переменной, когда независимые переменные фиксированы.

Регрессионный анализ широко используется для прогнозирования в сферах, связанных с машинным обучением. Регрессионный анализ также используется, чтобы понять, какие из независимых переменных связаны с зависимой переменной, и исследовать формы этих отношений. Регрессионный анализ также может использоваться, чтобы вывести причинно-следственные связи между независимыми и зависимыми переменными.

2 Критерии качества данных

При анализе качества данных важно учитывать:

- качество данных - многоаспектное понятие, включающее в себя множество критериев
- проблемы в критериях качества, таких как точность, могут быть как легко обнаружимы (например, орфографические ошибки), так и более трудные в распознавании (например, допустимые, но не правильные значения)
- полноту сложно оценить
- обнаружение непротиворечивости не всегда локализует ошибки

На основе изучения литературы ([2], [3], [5-10]) были выявлены основные критерии качества данных.

Величина	Источник	Изм.	Янв	Фев	Март
Средняя скорость ветра на высоте 10 м	NASA	%	04.08	4.2	4.19
Средняя скорость ветра на высоте 50 м	NSA	м/с	5.16	5.32	5.3
Повторяемость скорости ветра 0 - 2 м/с на высоте 50 м	NASA	м/с	8	7	9
Повторяемость скорости ветра 3 - 6 м/с на высоте 50 м	NASA	%	70	64	62

Таблица 1

2.1 Точность

Точность определяют как близость между значениями v и v’, где v – исследуется на корректность представления реального объекта.

Рассмотрим таблицу 1, если источником данных является NASA, то v’ = NASA – корректное значение, в то время, как v = NSA является некорректным. Существуют два вида точности – синтаксическая и семантическая.

2.1.1 Синтаксическая точность

Синтаксическая точность - близость значения v к элементам соответствующего домена D определения. В синтаксической точности мы не интересуемся сравнением v’ с истинным значением v; скорее мы интересуемся проверкой, является ли v кем-либо из значений в D, независимо от того, что это. Так, если v = GISMETEO, даже если v’ = NASA, v считают синтаксически корректным, поскольку GISMETEO - допустимое значение в домене источников информации. Синтаксическая точность измеряется посредством функций сравнения, которые оценивают расстояние между v и значениями в D.

2.1.2 Семантическая точность

Семантическая точность - близость значения v к истинному значению v’. Снова рассмотрим таблицу 1. “Обмен” единицами измерения в кортежах 1 и 3 является примером семантической ошибки точности.

2.2 Полнота

Интуитивно, полнота таблицы характеризует степень, которой таблица представляет соответствующий реальный мир.

В модели, предполагающей возможное наличие значений типа null, присутствие значения null имеет общее значение отсутствующего значения, то есть значение, которое существует в реальном мире, но по

некоторым причинам недоступно. Чтобы охарактеризовать полноту, важно понять, почему значение отсутствует. Действительно, значение может отсутствовать или потому что оно существует, но неизвестно, или потому что оно не существует вообще, или потому что оно может существовать, но неизвестно, существует ли оно на самом деле или нет.

В модели без нулевых значений, чтобы охарактеризовать полноту, мы должны представить понятие ссылочного отношения.

В модели с нулевыми значениями можно определить следующие уровни полноты:

- полнота значения характеризует полноту полей кортежа
- полнота кортежа характеризует полноту кортежа относительно значений всех его полей
- полнота атрибута определяет количество нулевых значений определенного атрибута в отношении
- полнота отношения определяет присутствие нулевых значений в целом отношении

2.3 Характеристики, связанные со временем

Важный аспект данных - их изменение и своевременное обновление. Основные связанные со временем критерии – это своевременность, употребительность и изменчивость.

Своевременность – критерий качества данных, измеряющий уровень доступности данных, когда процессы их требуют. Своевременность важна, поскольку нужные данные могут оказаться бесполезны из-за того, что они не приходят вовремя на запросы. Например, расписание занятий в университете не может быть своевременным, если появится в середине семестра.

Употребительность – критерий качества данных, показывающий, насколько данные соответствуют действительности в заданный момент времени.

Изменчивость характеризует частоту, с которой данные меняются во времени. Например, у стабильных данных, таких как даты рождения, изменчивость равна 0.

2.4 Непротиворечивость

Непротиворечивость характеризует противречия в данных. Предполагает наличие семантических правил, определенных по ряду элементов данных, где элементы могут быть кортежами реляционных таблиц или записей в файле. Со ссылкой на реляционную теорию ограничения целостности - инстанцирование таких семантических правил.

2.4.1 Ограничения целостности

Ограничения целостности - свойства, которые должны быть удовлетворены всеми экземплярами схемы базы данных. Несмотря на то, что ограничения целостности обычно определяются на схемах, они могут одновременно быть проверены и на определенном экземпляре схемы, который представляет расширение базы данных.

Вторая глава. Исследование методов улучшения качества данных

1 Входные данные

Для исследования местности, планируемой к застройке конкретным типом ВИЭ используется фиксированный набор климатических данных.

Для солнечных электростанций:

- Суммарная солнечная радиация на горизонтальную поверхность (КВтч/м^2/день)
- Суммарная солнечная радиация на наклонную поверхность (КВтч/м^2/день)
- Прямая солнечная радиация на нормальную поверхность (МДж/м^2)

Для ветровой энергетики:

- Средняя скорость ветра на высоте 10 м (м/с)
- Средняя скорость ветра на высоте 50 м NASA (м/с)
- Повторяемость скорости ветра 0 - 2 м/с на высоте 50 м (%)
- Повторяемость скорости ветра 3 - 6 м/с на высоте 50 м (%)
- Повторяемость скорости ветра 7 - 10 м/с на высоте 50 м (%)
- Повторяемость скорости ветра 11 - 14 м/с на высоте 50 м (%)
- Повторяемость скорости ветра 15 - 18 м/с на высоте 50 м (%)
- Повторяемость скорости ветра 19 - 25 м/с на высоте 50 м (%)

Это основные метеорологические параметры, остальные являются необязательными и больше нужны для проведения

аналитическо-экономических расчетов/обоснований строительства ВИЭ, поэтому они не были использованы в работе.

Величина	Источник	Изм.	Янв	Фев	Март	Апр	Май	Июнь	Июль	Авг	Сен	Окт	Нбр	Дек	Год
Средняя скорость ветра на высоте 10 м	NASA	м/с	04.08	4.2	4.19	4.35	4.12	3.99	3.83	04.01	4.27	4.33	04.09	3.88	4.11
Средняя скорость ветра на высоте 50 м	NASA	м/с	5.16	5.32	5.3	5.5	5.21	05.05	4.85	05.07	5.4	5.48	5.18	4.91	5.2
Повторяемость скорости ветра 0 - 2 м/с на высоте 50 м	NASA	%	8	7	9	7	10	12	15	12	9	8	9	9	10
Повторяемость скорости ветра 3 - 6 м/с на высоте 50 м	NASA	%	70	64	62	61	61	63	63	61	61	61	65	72	64
Повторяемость скорости ветра 7 - 10 м/с на высоте 50 м	NASA	%	21	28	28	31	28	24	22	27	29	30	25	18	26
Повторяемость скорости ветра 11 - 14 м/с на высоте 50 м	NASA	%	1	1	1	1	1	1	0	1	1	1	1	0	1
Повторяемость скорости ветра 15 - 18 м/с на высоте 50 м	NASA	%	0	0	0	0	0	0	0	0	0	0	0	0	0
Повторяемость скорости ветра 19 - 25 м/с на высоте 50 м	NASA	%	0	0	0	0	0	0	0	0	0	0	0	0	0
Суммарная солнечная радиация на горизонтальную поверхность	NASA	Втч/м2/ден	1.15	1.98	3.22	4.7	6	6.17	06.05	5.25	3.91	2.39	1.29	0.93	3.59
Суммарная солнечная радиация на наклонную поверхность	NASA	Втч/м2/ден	02.01	03.09	04.05	4.96	5.54	5.34	5.39	5.24	4.59	3.33	2.17	1.74	3.95
Прямая солнечная радиация на нормальную поверхность	NASA	МДж/м2	1.78	2.67	3.53	4.85	6	5.85	5.88	5.41	4.4	3.1	1.87	1.52	3.9

Рисунок 1. Пример полученных данных

Для исследования были взяты климатические данные по ветровой и солнечной энергетике по по Волгоградской и Ростовской области, а также по республике Калмыкии (использованы данные замеров в 100 точках с различными координатами). Источником является NASA SSE.

2 Оценка качества данных

Основными проблемами, характерными для представленных климатических данных, являются временные пробелы и аномальные пики значений. В данной работе представлены возможные способы повышения качества данных.

Величина	Источник	Изм.	Янв	Фев	Март	Апр	Май	Июнь	Июль	Авг	Сен	Окт	Нбр	Дек	Год
Средняя скорость ветра на высоте 10 м	NASA	м/с	04.08		4.19	4.35	4.12	3.99	3.83		4.27	4.33	04.09	3.88	4.11
Средняя скорость ветра на высоте 50 м	NASA	м/с	5.16	5.32	5.3	5.5	14.1	05.05	4.85	05.07	5.4	5.48	18.36	4.91	5.2
Повторяемость скорости ветра 0 - 2 м/с на высоте 50 м	NASA	%	8	7	9		10	12	15	12	9	8	9	9	10
Повторяемость скорости ветра 3 - 6 м/с на высоте 50 м	NASA	%	70	64	62	61	61	63	63	61	61	61		72	64
Повторяемость скорости ветра 7 - 10 м/с на высоте 50 м	NASA	%	21		28	31	28	24	22	78	29	30	25	18	26
Повторяемость скорости ветра 11 - 14 м/с на высоте 50 м	NASA	%	1	1	1	1	1		0	1	1	1	1	0	1
Повторяемость скорости ветра 15 - 18 м/с на высоте 50 м	NASA	%	0	0	0	0	0	0	0	0	0	0	0		0
Повторяемость скорости ветра 19 - 25 м/с на высоте 50 м	NASA	%	0	0	0	0	0	0	0	0	0		0	0	0
Суммарная солнечная радиация на горизонтальную поверхность	NASA	Втч/м2/де	1.15	1.98	3.22	4.7	6		06.05	5.25	3.91	2.39	1.29	0.93	3.59
Суммарная солнечная радиация на наклонную поверхность	NASA	Втч/м2/де	02.01	03.09		4.96	5.54	5.34	5.39	5.24	4.59	3.33	14	1.74	3.95
Прямая солнечная радиация на нормальную поверхность	NASA	МДж/м2	1.78	2.67	3.53	4.85	20	5.85	5.88	5.41		3.1	1.87	1.52	3.9

Рисунок 2. Пример данных с проблемами

2.1 Заполнение временных пробелов

Поскольку исходные данные не имеют пропусков, их необходимо внести случайным образом.

Существует четыре способа работы с пропущенным значением:

- удаление данных
- удаление переменной
- оценка значением
- прогнозирование

В рамках поставленной задачи первые два способа не являются корректными.

2.1.1 Оценка средним

Самый простой способ заполнения пропусков – заполнение средним значением. По некоторому атрибуту по всем известным значениям определяется среднее (так же возможен вариант с медианой, но значительных различий на представленных данных не обнаружено), далее этим значением заполняются все пропуски.

2.1.2 Повторение результата последнего наблюдения

В данном методе пропуски заполняются на основе последнего замера данных (или нескольких). Он считается наиболее эффективным для временных рядов, поэтому был выбран для использования на климатических данных. Важно учитывать, что при использовании метода повторения результата последнего наблюдения необходимо обрабатывать ситуацию, когда искомое значение содержится в первом наблюдении. В

этом случае наиболее эффективным является использование первого известного измерения.

2.1.3 Метод ближайших соседей

Метод ближайших соседей – классификатор, основанный на оценивании сходства. Классифицируемый объект относится к тому классу, которому принадлежат ближайшие к нему объекты обучающей выборки. Расстояние между объектами вычисляется с помощью метрики Евклида:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

В данной задаче он используется следующим образом: для пропущенного данного определяются ближайшие точки, а далее рассчитывается взвешенное среднее.

2.1.4 Rpart

Алгоритм rpart реализует рекурсивное разделение и строит дерево на основе исходных данных, на основе этого дерева производится вычисление пропущенного значения

2.1.5 Многомерная оценка оценка цепными уравнениями

Алгоритм использует способ оценки в два шага:

1. mice() для построения модели
2. complete() для генерации данных

Функция mice(df) создает несколько полных копий df, каждая со своей оценкой пропущенных данных. Функция complete() возвращает один или несколько наборов данных, набор по умолчанию будет первым.

2.1.6 Результаты

	MAE	MSE	RMSE
mean	3.246	16.772	4.095
last observe	2.648	10.574	3.252
kNN	2.003	7.916	2.81
rpart	1.421	3.988	1.997
mice	0.730	3.124	1.768

2.2 Сглаживание аномальных пиков значений

Аналогично предыдущему пункту, перед использованием алгоритмов необходимо внести в исходные данные аномальные значения. Но в отличие от пропусков, это задача требует большей осмысленности, поскольку на основе представленных значений необходимо определить величины, которые можно будет отнести к аномальным.

Для сглаживания были выбраны два алгоритма: сглаживающий сплайн и разложение в сумму синусов.

После внесения аномальных значений средняя абсолютная ошибка составила 14.852.

Исследованные методы сглаживания можно разбить на две группы:

- сглаживание с помощью аппроксимации исходных данных с помощью функции (сглаживающий сплайн и разложение в сумму синусов)
- сглаживание с помощью использования данных соседних точек (метод скользящего среднего и взвешенная локальная регрессия)

2.2.1 Сглаживающий сплайн

Хотя сглаживающий сплайн и относят к непараметрическим моделям, тем не менее он содержит задаваемый пользователем параметр. Сглаживающий сплайн определяется как сплайн, который минимизирует следующий функционал, зависящий также и от некоторого параметра *p*.

$$I(s;p) = p\sum_{k=1}^{n} w_k (y_k - s(x_k))^2 + (1-p)\int_{x_1}^{x_n}\left(\frac{d^2 s(x)}{dx^2}\right)^2 dx$$ [4]

где

$(x_k,y_k)_k=1,2,...,n$ - приближаемые данные;

w_k - веса данных (если они не были заданы, то принимаются равными единице);

p - сглаживающий параметр, изменяющийся от 0 до 1, который определяет кривизну получающегося сплайна.

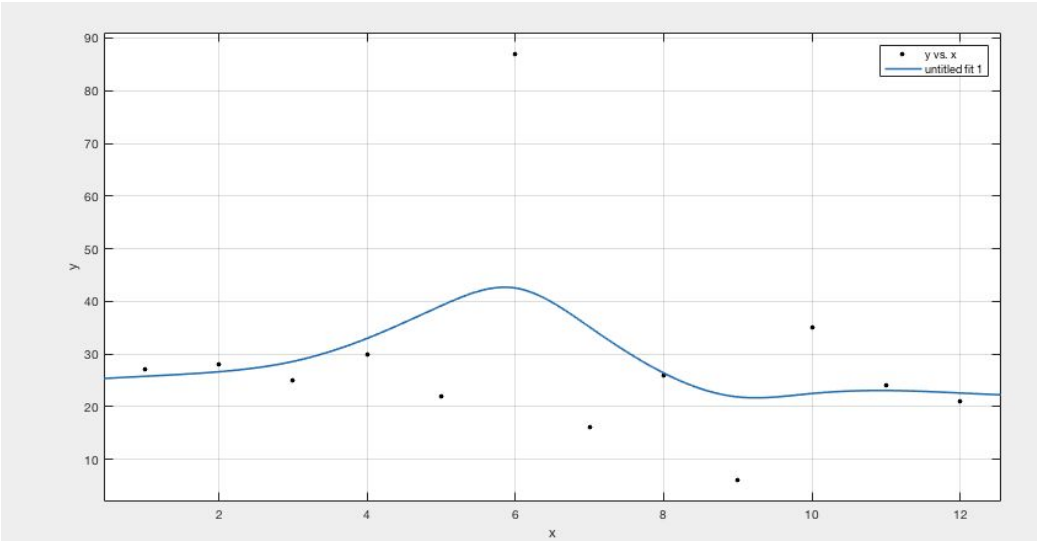

Рисунок 3. Применение сглаживающего сплайна

[4] URL: http://matlab.exponenta.ru/curvefitting/3_5.php (Дата обращения: 10.05.2017)

Средняя абсолютная ошибка при использовании данного метода составила 5.990.

2.2.2 Разложение в сумму синусов

Данный метод использует аппроксимацию исходных данных с помощью модели вида

$$a_1 \sin(b_1 x + c_1)$$
$$a_1 \sin(b_1 x + c_1) + a_2 \sin(b_2 x + c_2)$$
$$\vdots$$
$$a_1 \sin(b_1 x + c_1) + a_2 \sin(b_2 x + c_2) + a_8 \sin(b_8 x + c_8)$$

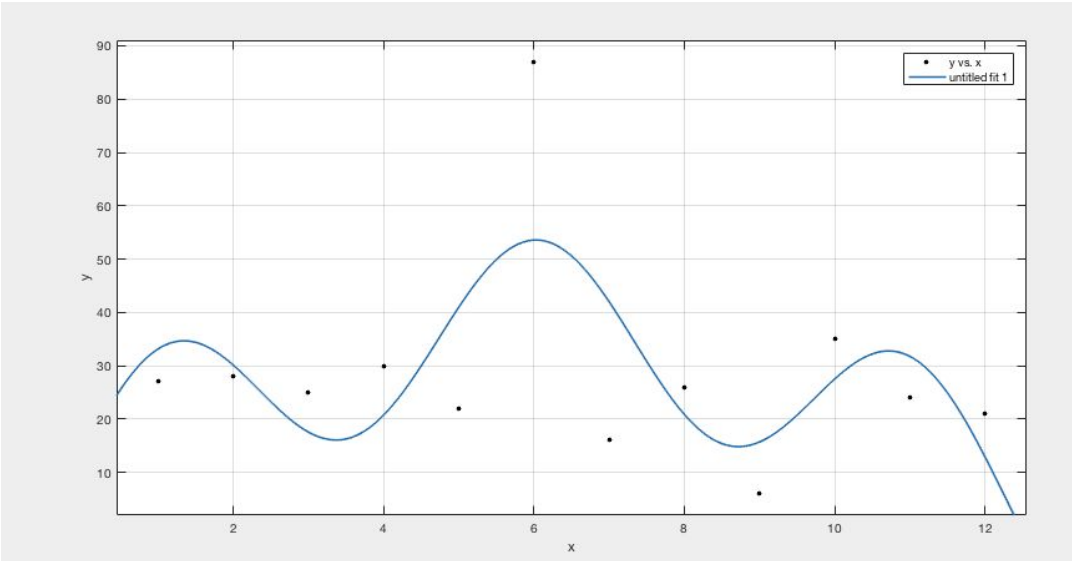

Рисунок 4. Применение разложения в сумму синусов

Средняя абсолютная ошибка при использовании данного метода составила 8.475.

2.2.3 Метод скользящего среднего

В методе скользящего среднего исходные сглаживаются по следующему правилу

$$ys_i = \frac{1}{2N+1} \sum_{k=-N}^{N} y(i+k),$$

где *2N + 1* – число точек, выбираемых для сглаживания. Данные, расположенные близко к границам отрезка, не сглаживаются, так как не хватает точек справа или слева от текущей, в которой в данный момент производится сглаживание.

Средняя абсолютная ошибка при использовании данного метода составила 4.337.

2.2.4 Взвешенная локальная регрессия

В данном методе для каждого сглаживаемого значения данных выбирается набор из фиксированного числа рядом расположенных точек, каждой из которых назначается вес по следующей формуле

$$w_i = \left(1 - \left|\frac{x_k - x_i}{d(x_k)}\right|^3\right)^3,$$

где $d(x_k)$ - расстояние от x_k до наиболее удаленной точки из набора. [5]

После выбора весов сглаженное значение находится при помощи локальной взвешенной регрессии с определенными выше весами:

- Lowess – линейная регрессия
- Loess – квадратичная регрессия

Средняя абсолютная ошибка при использовании данного метода составила 5.743(Lowess) и 4.176(Loess).

[5] URL: http://matlab.exponenta.ru/curvefitting/3_10.php#02 (Дата обращения: 10.05.2017)

Результаты

В рамках данной работы были изучены и применены к реальным климатическим данным ряд методов улучшения качества данных, решающие проблемы пропущенных значений и аномальных выбросов. На основании проведенных измерений можно утверждать, что для заполнения пропусков большую точность показали методы, основанные на прогнозировании. Метод повторения результата последнего измерения также показал достаточно высокую точность, поскольку исходные климатические данные представляют собой подобие временных рядов. По этой же причине для сглаживания аномальных значений наиболее оптимальными оказались метод скользящего среднего и взвешенная локальная регрессия, которые сглаживают значение в точке, основываясь на соседних значениях.

В ходе выполнения работы были выполнены следующие задачи:

- Изучено понятие и критерии качества данных
- Изучены и получены данные, используемые для исследования местности, планируемой к застройке ВИЭ
- Изучены алгоритмы заполнения пропущенных значений и выявлен оптимальный для исходных данных (mice)
- Изучены алгоритмы сглаживания аномальных значений и выявлен оптимальный для исходных данных (Взвешенная локальная регрессия)

Таким образом, поставленная задача была решена.

Список литературы

[1] Барсегеян А. А., Куприянов М. С. Методы и модели анализа данных: OLAP и Data Mining СПб.: БХВ-Петербург 2004.

[2] Batini C. et al. Methodologies for data quality assessment and improvement //ACM computing surveys (CSUR). – 2009. – Т. 41. – №. 3. – С. 16.

[3] Sadiq S. Handbook of data quality //New York: Springer. do i. – 2013. – Т. 10. – С. 978-3.

[4] Сирота А. А. Методы и алгоритмы анализа данных и их моделирование в MATLAB СПб.: БХВ-Петербург 2017

[5] Redman T. C., Blanton A. Data quality for the information age. – Artech House, Inc., 1997.

[6] Evans P. Scaling and assessment of data quality //Acta Crystallographica Section D: Biological Crystallography. – 2006. – Т. 62. – №. 1. – С. 72-82.

[7] Strong D. M., Lee Y. W., Wang R. Y. Data quality in context //Communications of the ACM. – 1997. – Т. 40. – №. 5. – С. 103-110.

[8] Wand Y., Wang R. Y. Anchoring data quality dimensions in ontological foundations //Communications of the ACM. – 1996. – Т. 39. – №. 11. – С. 86-95.

[9] Wang R. Y., Strong D. M. Beyond accuracy: What data quality means to data consumers //Journal of management information systems. – 1996. – Т. 12. – №. 4. – С. 5-33.

[10] Sattler K. U. Data Quality Dimensions //Encyclopedia of Database Systems. – Springer US, 2009. – С. 612-615.

[11] Kaufman L., Rousseeuw P. J. Finding groups in data: an introduction to cluster analysis. – John Wiley & Sons, 2009. – Т. 344.

[12] Бальзанников М. И., Елистратов В. В. Возобновляемые источники энергии //Аспекты комплексного использования. Самара: Офорт. – 2008.

[13] БАЗЕ Э., НАСОСА С. Э. И. Т. ВОЗОБНОВЛЯЕМЫЕ ИСТОЧНИКИ ЭНЕРГИИ //РАДИОЭЛЕКТРОНИКА, ЭЛЕКТРОТЕХНИКА И ЭНЕРГЕТИКА: Двадцать первая Междунар. науч.-техн. конф. студентов и аспирантов: Тез. докл. В 4 т. Т. 4. М.: Издательский дом МЭИ, 2015.—303 с. – С. 245.

[14] Herzog A. V., Lipman T. E., Kammen D. M. Renewable energy sources //Encyclopedia of Life Support Systems (EOLSS). Forerunner Volume-'Perspectives and Overview of Life Support Systems and Sustainable Development. – 2001.

[15] Boyle G. et al. Renewable energy: power for a sustainable future. – OXFORD university press, 1997.

[16] Liserre M., Sauter T., Hung J. Y. Future energy systems: Integrating renewable energy sources into the smart power grid through industrial electronics //IEEE industrial electronics magazine. – 2010. – Т. 4. – №. 1. – С. 18-37.

[17] Demirbas A. Potential applications of renewable energy sources, biomass combustion problems in boiler power systems and combustion related environmental issues //Progress in energy and combustion science. – 2005. – Т. 31. – №. 2. – С. 171-192.